New Phytologist

Research

Conceptual analysis of methods applied to assessment of diversity within and distance between populations with asexual or mixed mode of reproduction

Evsey Kosman[1] and Kurt J. Leonard[2]

[1]Institute for Cereal Crops Improvement (ICCI), The George S. Wise Faculty for Life Sciences Tel Aviv University, Tel Aviv 69978, Israel; [2]Department of Plant Pathology, University of Minnesota, St Paul, MN 55108, USA

Author for correspondence:
E. Kosman
Tel: +972 (0)3 6407857
Fax: +972 (0)3 6407857
Email: kosman@post.tau.ac.il

Received: *16 October 2006*
Accepted: *9 January 2007*

Summary

• Measures of diversity within populations, and distance between populations, are compared for organisms with an asexual or mixed mode of reproduction. Examples are drawn from studies of plant pathogenic fungi based on binary traits including presence/absence of DNA bands or virulence/avirulence to differential hosts.
• Commonly used measures of population diversity or genetic distance consider either genotype frequencies or allele frequencies. Kosman's diversity and distance measures are the most suitable for populations with an asexual or mixed mode of reproduction, because by considering genetic patterns of all individuals they take into account not just the genotype frequencies but also the genetic similarities between genotypes in the populations.
• The Kosman distance and diversity measures for populations can be calculated using different measures of dissimilarity between individuals (the simple mismatch, Jaccard and Dice coefficients of dissimilarity). Kosman's distances based on the simple mismatch and Jaccard dissimilarities are metrics.
• Comparisons of diversity indices for hypothetical examples as well as for actual data sets are presented to demonstrate that inferences from diversity analysis of populations can be driven by techniques of diversity and distance assessments and not only data driven.

Key words: assignment problem, clustering, genetic diversity, Kosman indices, plant pathogens, population genetics.

New Phytologist (2007) **174**: 683–696

doi: 10.1111/j.1469-8137.2007.02031.x

Introduction

In large, randomly mating populations that are in Hardy–Weinberg equilibrium with little or no linkage disequilibrium, there is a predictable relationship between allele frequencies at any given series of marker loci and the observed frequencies of genotypes defined by those loci. For such populations, measurements of genetic diversity within or genetic distance between populations can be based on allele frequencies without the need to compare genetic similarities between individual genotypes that make up those populations. Conversely, species that reproduce asexually to a significant extent generally exhibit marked linkage disequilibria for alleles between marker loci. That is, the allele frequencies in asexual populations generally do not relate directly to frequencies of specific genotypes within the populations. The observed genotype frequencies may deviate strongly above or below their expected frequencies calculated from the products of the observed frequencies of individual alleles in those genotypes. For predominantly asexual populations, it is reasonable to assume that genotypes that are identical at all marker loci share a common ancestor and that genotypes that differ in alleles at only one or a few marker loci are more closely related than genotypes that differ at many loci. To take

this into account, unconventional measures of genetic diversity and genetic distance may be needed when populations of predominantly asexual species are compared.

Many organisms, including plant pathogenic fungi, undergo asexual or a mixed mode of reproduction. Hundreds of published research studies have addressed the diversity and population structure of plant pathogenic fungi, because of the social significance and economic importance of plant diseases. Therefore, plant pathogen populations can be considered among the most suitable models for comprehensive analysis of methods used to study populations of asexually reproducing organisms.

Various indices of diversity within and distance between populations are used for comparative analyses of plant pathogen populations. In many studies, however, the specific diversity index employed in the analyses appears to have been arbitrarily chosen without critical attention to its appropriateness for the type of data to be analysed. Although some recommendations for assays of genetic variation were proposed by Grünwald *et al.* (2003) and limited comparative analyses using different indices have been attempted (Kosman, 1996, 2003; Manisterski *et al.*, 2000), there has not been a comprehensive analysis of the merits of different diversity indices and the consequences of uncritical choices of indices in studies of the types of populations typically encountered with plant pathogenic fungi. In examining these issues, we have found that diversity indices developed for species diversity in ecosystems or for population diversity of organisms that reproduce only sexually in randomly mating populations are based on assumptions that may not apply in comparisons of populations of fungal species that undergo asexual reproduction in addition to or in place of sexual reproduction. Furthermore, we found that employing an inappropriate diversity or genetic distance estimator can lead to misleading conclusions in studies of diversity in fungal populations.

Distance based clustering methods proceed by calculating a matrix of distances between populations followed by any convenient graphical representation. The following distance measures have been most frequently used: the Rogers distance (Rogers, 1972), Nei's genetic distance (Nei, 1972, 1978), Nei's coefficient of differentiation (Nei, 1973), the average differences between populations (McCain *et al.*, 1992), Kosman's distance between populations (Kosman, 1996) and the mean character difference (Sneath & Sokal, 1973; Long *et al.*, 1998; Manisterski *et al.*, 2000). The structure of clusters based on these different measures will not necessarily be the same for a given data set. Therefore, the choice of which measure to use should be justified by the type of data obtained. The main objectives of this manuscript are to explore potential problems, to demonstrate how diversity analysis and clustering results can depend on the approach used, and to consider the biological rationale for different approaches. We show that not all diversity and distance measures adequately represent basic properties of diversity within and distance between populations, respectively. We use quite simple examples of artificial and actual binary data to demonstrate that relationship between populations based on diversity within and between them can be driven by the technique of diversity assessment and not only data driven. The results can easily be extended to multistate discrete characters (data in any nominal scale). However, to simplify explanations multistate characters are not considered in the manuscript. We expect to address the concept of sampling variance associated with an estimate of population statistic in a future study.

Types of diversity indices

The following provides a general analysis of relationships between different diversity indices and explains some properties of these indices. Application of these indices in a few abstract examples as well as actual data sets are included to illustrate how properties of the indices may affect results and interpretations of the analyses. For convenient reference, all coefficients and indices mentioned and discussed in the text are presented together in Appendix A. In the examples, we express molecular and virulence markers data in binary characters for presence/absence of DNA bands or virulence/avirulence of plant pathogens to differential hosts. Virulence markers used in analyses of genetic diversity in plant pathogens typically occur in polymorphic gene-for-gene relationships in which each major gene for resistance in the host plant is effective only if it is matched by a corresponding gene for avirulence in the pathogen. The resistance gene is ineffective if the pathogen is homozygous for the virulence gene at the corresponding locus. For diversity analyses, a virulent or avirulent reaction may be regarded as equivalent to the presence or absence of a band at a specific location in a DNA gel.

Designations and dissimilarity between individuals

Consider a sample from population P and two samples collected from two populations P_1 and P_2, which consist of the same number n of individuals tested on k differentiating factors and represented by binary patterns of 1 for positive response (virulence, presence of a band) and 0 for negative (avirulence, absence of a band). If the numbers of individuals in P_1 and P_2 are different, then two bootstrapped samples of the same size, n, from P_1 and P_2 could be considered. We denote by q_i, q_{1i} and q_{2i} the frequencies of appearance of 1 at the ith differentiating factor for populations P, P_1 and P_2, respectively. For example, if the differentiating factors comprise a typical set of differential host lines used in virulence tests, q_i would be the frequency of virulence in population P on the ith differential line. The frequencies of genotype r in populations P, P_1 and P_2 are denoted by p_r, p_{1r} and p_{2r}, respectively. We denote by ρ any coefficient of dissimilarity between two individuals x_1 and x_2, $\rho(x_1, x_2)$, and we compare results from use of the commonly used dissimilarity measures: simple mismatch coefficient, m,

Table 1 Diversity and distance measures: basis for calculation, mathematical relationships and applicability

Basis for calculation of estimators	Diversity within population	Distance between populations	Applicability
Dissimilarity between individuals	Average difference (*ADW*)[ab]	Distance of average differences (*DAD*)[c]	Sexual reproduction random mating linkage equilibrium genotypic diversity[a]
	Kosman diversity (*KW*)	Kosman distance (*KB*)[d]	Asexual reproduction non-random mating linkage disequilibrium genotypic diversity[d]
Frequencies of differentiating factors	Nei gene diversity (H_S)[b] Kosman index (*K*)	Mean character differences (*MCD*) Nei standard genetic distance (*N*) Nei minimum genetic distance (N_M)[c] Nei coefficient of differentiation (G_{ST})	Sexual reproduction random mating linkage equilibrium
Frequencies of genotypes	Simpson index (*Si*)[a] Stoddart index (*St*) Shannon evenness index (*E*) Shannon normalized index (*Sh*)	Rogers distance (*R*)[d]	Genotypic diversity species diversity

[a]If genetically different individuals are considered equally distinct (Eqn 6), the average difference within a population (*ADW*) is equivalent to the Simpson index (*Si*) of genotypic diversity (Eqn 8).
[b]Nei's measure of gene diversity per locus (H_S) equals the average difference within a population (*ADW*) with respect to the simple mismatch dissimilarity between individuals (Eqn 4), although H_S was originally defined in terms of frequencies of alleles (differentiating factors).
[c]The distance of average differences (*DAD*) with respect to the simple mismatch coefficient of dissimilarity between individuals equals Nei's minimum genetic distance (N_M).
[d]If genetically different individuals are considered equally distinct (Eqn 6), the Kosman distance (*KB*) is equivalent to the Rogers index (*R*) of genotypic distance (Eqn 7).

Euclidean distance, *e*, Jaccard, *j*, and Dice, *d*, coefficients of dissimilarity (see Appendix A1). A recently developed measure of dissimilarity between diploid organisms tested with codominant molecular markers (Kosman & Leonard, 2005) could also be considered.

Classification of diversity indices

Measures of diversity within and distance between populations may be initially divided into three groups according to sources of information used for their calculation (Table 1). These are indices based on: (1) dissimilarity between individuals; (2) frequencies of appearance of differentiating factors (alleles or DNA bands); and (3) frequencies of individuals' genotypes determined according to a set of differentiating factors. The first group, based on dissimilarity between individuals, includes the average difference within (*ADW*) and between (*ADB*) populations (McCain *et al.*, 1992), Kosman's distance between (*KB*) and diversity within (*KW*) populations (Kosman, 1996; Manisterski *et al.*, 2000; Schachtel & Kosman, 2002) and the Müller index (*Mu*) of diversity (Müller *et al.*, 1996). The second group, based on factor (allele) frequencies, includes Nei's standard genetic distance (*N*) and Nei's minimum genetic distance (N_M) (Nei, 1972, 1978), mean character difference (*MCD*) (Sneath & Sokal, 1973; Long *et al.*, 1998; Manisterski *et al.*, 2000), Nei's coefficient of differentiation (G_{ST}) and Nei's measure of the average gene diversity per locus (H_S) (Nei, 1973), and Kosman's index (*K*) (Manisterski *et al.*, 2000). The third group, based on genotype frequencies, includes Rogers distance (*R*) (Rogers, 1972), Simpson's diversity index (*Si*) (Simpson, 1949), Stoddart's index (*St*) (Stoddart, 1983; Stoddart & Taylor, 1988), Shannon's entropy (*Sh*) (Shannon & Weaver, 1949) and the Shannon evenness parameter (*E*) (Sheldon, 1969).

This three-way division of the indices is not absolute. For example, Manisterski *et al.* (2000) showed that Müller's mean dissimilarity *Mu* could be considered an index from the second group, because it may be expressed by the virulence frequencies. Moreover, the Müller index can be considered as the correction of Nei's measure of the average gene diversity per locus H_S for small samples (Kosman, 2003) due to the following equalities:

$$\begin{aligned} Mu &= \frac{2n}{k(n-1)} \cdot \sum_{i=1}^{k} q_i(1-q_i) \\ &= \frac{n}{n-1} \cdot \frac{1}{k} \cdot \sum_{i=1}^{k} [1 - q_i^2 - (1-q_i)^2] \\ &= \frac{n}{n-1} \cdot H_S. \end{aligned}$$

Thus, Nei's diversity $H_S = (1 - 1/n)Mu$, and it may be considered an index from the first group. In addition, we will further show that the Rogers distance and the Simpson

diversity from the third group and the Nei genetic distance and Nei's minimum genetic distance from the second group may also be represented as indices based on dissimilarity between individuals (i.e. first group).

Indices of average differences

The indices of average difference within and between populations can be defined for any measure of dissimilarity between individuals ρ. We denote as ADW_ρ and ADB_ρ the average differences within and between populations with respect to measure of dissimilarity ρ:

$$ADW_\rho(P) = \frac{1}{n^2} \cdot \sum_{i,j=1}^{n} \rho(x_i,x_j) \quad \text{and}$$
$$ADB_\rho(P_1,P_2) = \frac{1}{n^2} \cdot \sum_{i,j=1}^{n} \rho(x_{1i},x_{2j}) \qquad \text{Eqn 1}$$

(x_i, x_{1i} and x_{2i} are individuals from populations P, P_1 and P_2, respectively). Average differences within and between populations calculated by the Jaccard (Adhikari *et al.*, 1999) and the Dice (Table 3 in Kolmer & Liu, 2000) measures, the Euclidean distance and the simple mismatch coefficient (Table 2 in Kolmer & Liu, 2000) may characterize populations in qualitatively distinct ways, because the sums of dissimilarities between multiple pairs of individuals will differ when described by the different measures. For instance, examples can be constructed in which

$$j(y_1,y_3) + j(y_1,y_4) < j(y_2,y_3) + j(y_2,y_4)$$

but $\quad d(y_1,y_3) + d(y_1,y_4) > d(y_2,y_3) + d(y_2,y_4),$

where $j(y_1,y_3)$ is the Jaccard dissimilarity between individuals y_1 and y_3, $d(y_1,y_3)$ is the Dice dissimilarity between individuals y_1 and y_3, etc. (data not shown). For dominant markers in diploid or dikaryotic fungi, the simple mismatch dissimilarity measure is superior to either the Dice or Jaccard dissimilarity measures for analyses of genetic diversity within populations or genetic distance between populations of the same species. For dominant markers, both the Dice and Jaccard measures ignore similarities between individuals that share the absence of a dominant trait, whereas both shared presence or shared absence of the dominant trait are taken into account by the simple match coefficient (Kosman & Leonard, 2005).

The index of average difference within population ADW_ρ may be used as measure of diversity within a population which ranges from 0 to 1 if $0 \le \rho \le 1$, and $ADW_\rho(P) = 0$ if and only if population P consists of individuals of identical genotype. By contrast, the index of average difference between populations ADB_ρ cannot be used as measure of distance between populations, because it generally distinguishes between two identical populations: $ADB_\rho(P,P) = ADW_\rho(P) > 0$ if P includes at least two individuals of different genotype. The distance of average differences between populations DAD_ρ could be proposed as a correction of ADB_ρ index:

$$DAD_\rho(P_1,P_2) = ADB_\rho(P_1,P_2) - \frac{ADW_\rho(P_1) + ADW_\rho(P_2)}{2} \qquad \text{Eqn 2}$$

The problem is that the DAD_ρ index may sometimes take negative values for certain dissimilarity measures ρ (E. Kosman, unpublished). One can prove that the distance of average differences with respect to the simple mismatch coefficient DAD_m equals Nei's minimum genetic distance N_M (Nei, 1972). It is always nonnegative, $DAD_m(P_1,P_2) \ge 0$, and $P_1 = P_2$ implies $DAD_\rho(P_1,P_2) = 0$. Therefore, DAD_m for the simple mismatch coefficient could be used as measure of distance between populations.

Assessments of distance between populations using among populations diversity corrected for diversities within populations were also considered in earlier studies (for Dice dissimilarity, Lynch, 1990; Lynch & Milligan, 1994). However, we did not find any thorough justification of the validity of the methods proposed. Therefore, we would not recommend application of the distance of average differences with respect to the Jaccard or Dice dissimilarity because there is no proof yet that values of DAD_j and DAD_d are always nonnegative.

One can show that the Nei genetic distance between two populations may be represented in the form

$$N(P_1,P_2) = -\ln\frac{1 - ADB_m(P_1,P_2)}{\sqrt{(1 - ADW_m(P_1))(1 - ADW_m(P_2))}} \qquad \text{Eqn 3}$$

This means that the Nei distance can be considered an index from the first group, and that it is a function of the simple mismatch dissimilarities between individuals. The first approximation of the Nei distance is the distance of average differences with respect to the simple mismatch coefficient: $N(P_1,P_2) \approx ADB_m(P_1,P_2) - (ADW_m(P_1) + ADW_m(P_2))/2 = DAD_m(P_1,P_2)$. This approximation is more accurate when values of $ADB_m(P_1,P_2)$, $ADW_m(P_1)$ and $ADW_m(P_2)$ are closer to zero (P_1 and P_2 are closely related and quite homogeneous populations).

As was proved in (Kosman, 2003), Nei's measure of the average gene diversity per locus H_S and the Müller diversity *Mu* within population *P* can be expressed by the index of average differences within population ADW_m with respect to the simple mismatch dissimilarity:

$$H_S(P) = ADW_m(P) \quad \text{and} \quad Mu(P) = \frac{n}{n-1} ADW_m(P) \qquad \text{Eqn 4}$$

Thus, the Nei diversity H_S and ADW_m index are mathematically identical, calculation of both estimators for a

data set (for instance, Table 1 in Gale *et al.*, 2002) is a tautology, and therefore similarity of conclusions from these two measures does not show robustness of the results.

Kosman assignment based indices

The Kosman distance *KB* (Kosman, 1996) between two populations P_1 and P_2 of n individuals is defined as follows. To each individual from P_1 an individual from P_2 is matched so as to minimize the sum of dissimilarities between n corresponding pairs of individuals (optimal matching of individuals with similar genotypes). Finding the best matches is known as the 'assignment problem', and Kosman's distance is obtained by dividing the calculated minimum value of the sum of dissimilarities between matched pairs of individuals $Ass_{\min}(P_1,P_2)$ by the number n of matched pairs. If the numbers of individuals per population are not equal, two bootstrapped samples of the same size can be derived from the two populations for comparison. There are $n!$ possible ways to match pairs of individuals between populations P_1 and P_2. When n is large, the number of possibilities increases very steeply. To achieve a workable solution, the 'assignment problem' algorithm provides an approach that eliminates a certain proportion of possibilities that have no chance of providing a minimum overall dissimilarity value. Even so, it is necessary to analyse great numbers of remaining possible combinations of matched pairs to find the combination that gives the minimum dissimilarity value.

The following example makes more tangible the method of calculation of Kosman distance.

Example 1 Consider samples of three individuals each from populations P_1 and P_2 tested on six differentials.

	P_1							P_2						
	1	2	3	4	5	6		1	2	3	4	5	6
a_1	1	1	1	1	0	0	b_1	1	1	1	1	1	0
a_2	1	1	0	1	1	1	b_2	0	1	1	0	0	0
a_3	0	1	1	1	1	1	b_3	0	0	1	0	0	0

According to the simple mismatch coefficient m, dissimilarities between the individuals from these samples are represented by the following matrix:

m	a_1	a_2	a_3
b_1	1/6	2/6	2/6
b_2	2/6	5/6	3/6
b_3	3/6	6/6	4/6

In general, there are 6 ($n!$ for $n = 3$) possibilities to match individuals from P_1 and P_2. The following two groups of matched pairs result in minimum of the sum of dissimilarities between corresponding pairs of individuals: $A = \{(a_1,b_3),(a_2,b_1),(a_3,b_2)\}$ and $B = \{(a_1,b_2),(a_2,b_1),(a_3,b_3)\}$. Then $Ass_{\min}(P_1,P_2) = m(a_1,b_3) + m(a_2,b_1) + m(a_3,b_2) = m(a_1,b_2) + m(a_2,b_1) + m(a_3,b_3) = 8/6$. The most similar pair of individuals (a_1,b_1) does not belong to any group of matched pairs, A and B. Moreover, pairs (a_1,b_3) from A and (a_3,b_3) from B are the most dissimilar possible pairs for individuals a_1 and a_3, respectively. Therefore, an algorithm for solving the 'assignment problem' requires analysing a great number of possible combinations of matched pairs, but not all of them ($n!$), and selecting at least one that gives the minimum overall dissimilarity value. Finally, the Kosman distance between populations P_1 and P_2 equals $Ass_{\min}(P_1,P_2)/n = 8/18 = 4/9$.

The Kosman diversity *KW* (Kosman, 1996) within population P of n individuals is defined as follows. Individuals from P are matched to make up n pairs so as to maximize the sum of dissimilarities between the corresponding pairs (optimal matching of individuals with dissimilar genotypes). Finding such matches can be realized by the solution of the appropriate 'assignment problem', and Kosman's diversity is determined by dividing the obtained maximum value of the sum of dissimilarities between matched pairs of individuals $Ass_{\max}(P,P)$ by the number n of matched pairs.

Kosman's diversity within and distance between populations can be defined for different measures of dissimilarity between individuals. We will denote as KB_ρ and KW_ρ the Kosman diversities between and within populations with respect to dissimilarity ρ:

$$KB_\rho(P_1,P_2) = \frac{1}{n} Ass^\rho_{\min}(P_1,P_2) \quad \text{and}$$
$$KW_\rho(P) = \frac{1}{n} Ass^\rho_{\max}(P,P) \qquad \text{Eqn 5}$$

Kosman's distance between populations KB_ρ is always nonnegative contrary to the distance of average differences DAD_ρ which may score negative values for some dissimilarity measures ρ. One can prove that if dissimilarity ρ between individuals is metric then the Kosman distance between populations KB_ρ is also metric (see Appendix B). This property of Kosman's distance may be important for some applications.

Let us show that the Rogers distance between two populations P_1 and P_2 and the Simpson diversity index can be represented as indices based on dissimilarity between individuals. Discrete measure of dissimilarity (discrete metrics) between individuals x_1 and x_2 is defined as follows

$$\delta(x_1,x_2) = \begin{cases} 1, & \text{if } x_1 \text{ and } x_2 \text{ are of different genotypes} \\ 0, & \text{if } x_1 \text{ and } x_2 \text{ are of same genotypes} \end{cases} \qquad \text{Eqn 6}$$

Let the numbers of individuals of genotype r from populations P_1 and P_2 be n_{1r} and n_{2r}, respectively, and s is the total number

of genotypes in both populations of n individuals. Then for a single differentiating factor the Kosman distance between populations P_1 and P_2 with respect to dissimilarity δ equals the Rogers distance between these populations:

$$KB_{\delta}(P_1,P_2) = \frac{Ass^{\delta}_{\min}(P_1,P_2)}{n} = \frac{1}{n}\cdot\frac{1}{2}\sum_{r=1}^{s}\left|n_{1r}-n_{2r}\right| = \frac{1}{2}\sum_{r=1}^{s}\left|p_{1r}-p_{2r}\right| = R(P_1,P_2) \quad \text{Eqn 7}$$

This means that with more than one differentiating factor the Kosman distance with respect to the simple mismatch coefficient $KB_m(P_1,P_2)$ (Kosman, 1996) can be considered as a generalization of the Rogers distance, in which the measure of similarity between different genotypes is taken into account. One can prove that $\min\{n_{1r},n_{2r}\}$ individuals from population P_2 are matched to the individuals of genotype r from population P_1 by the solution of the corresponding 'assignment problem' $Ass_{\min}(P_1,P_2)$. Thus, the Kosman distance KB_m takes into account both the genotypic structure of populations and the measure of similarity between different genotypes.

Let n_r be the number of individuals of genotype r from population P of n individuals, and s is the total number of genotypes in this population. Then the average difference within population P with respect to dissimilarity δ equals the Simpson diversity index within this population:

$$ADW_{\delta}(P) = \frac{1}{n^2}\sum_{i,j=1}^{n}\delta(x_i,x_j) = \frac{1}{n^2}\sum_{r=1}^{s}n_r(n-n_r) = \sum_{r=1}^{s}p_r(1-p_r) = 1-\sum_{r=1}^{s}p_r^2 = Si(P) \quad \text{Eqn 8}$$

Thus, the Rogers distance and the Simpson diversity, which were originally defined as functions of frequencies of genotypes, can be considered as indices based on dissimilarities of individuals with respect to discrete metrics.

Genotype based indices

The Shannon entropy $-\sum_{r=1}^{s}p_r\ln p_r$ (Shannon & Weaver, 1949) measures diversity within population on the basis of frequencies of genotypes p_r, $r = 1, 2, \ldots, s$. Its value ranges between 0 for a single genotype and $\ln s$ when genotypes are evenly distributed ($p_r = 1/s$ for all r). If the number s of genotypes increases then the Shannon entropy may tend to infinity. For convenience, it is possible to normalize this index in order to make it quantitatively comparable with other diversity measures, which vary between 0 and 1. There are two ways to normalize the Shannon entropy. The first one was proposed by Sheldon (1969) and implies division of the Shannon entropy by its maximum value $\ln s$ for a sample which comprises the same number s of genotypes. This is the so-called evenness parameter:

$$E(P) = -\frac{\sum_{r=1}^{s}p_r\ln p_r}{\ln s}.$$

The second way implies division of the Shannon entropy by its maximum value $\ln n$ for a sample that consists of the same number n of individuals (in this case n is also the maximum possible number of genotypes for such a sample) (Goodwin *et al.*, 1992; Andrivon & de Vallavieille-Pope, 1995). We will call this measure of diversity within populations the Shannon normalized index:

$$Sh(P) = -\frac{\sum_{r=1}^{s}p_r\ln p_r}{\ln n}.$$

Both evenness and richness aspects of genotypic diversity are taken into account by this last index because it can be represented as the product of the evenness and richness parameters of the population as follows:

$$Sh(P) = -\frac{\sum_{r=1}^{s}p_r\ln p_r}{\ln s}\cdot\frac{\ln s}{\ln n} = E(P)\cdot\frac{\ln s}{\ln n} \quad \text{Eqn 9}$$

where the richness parameter $\ln s/\ln n$ reflects relative abundance of the sample. Hence, the Shannon normalized index is more informative than the evenness parameter, and therefore we would recommend to use this index for measuring genotypic diversity. We agree only in part with the conclusion of Grünwald *et al.* (2003) that scaling the Shannon entropy by sample size ($\ln n$) should be avoided. This conclusion is right only in cases in which the maximum theoretically possible number of genotypes that can be detected by the technique used to assay for variation is less than the sample size. For example, if 1000 isolates are tested on eight binary differentiating factors (virulence/avirulence, presence/absence of DNA bands) then the maximum theoretically possible number of genotypes equals $2^8 = 256$. In this case (similar to simulation parameters considered by Grünwald *et al.* (2003): 1000 for sample size and 200 for number of genotypes) scaling by sample size may distort actual diversity assessments because the richness component $\ln s/\ln n$ of the Shannon normalized index in equation 9 never reaches its maximum theoretically possible value 1 ($\ln s/\ln n \leq \ln 256/\ln 1000 \approx 0.803$). However, a number of differentiating features used in regular diversity analyses of plant pathogen populations typically varies from 10 to 30 differential hosts for virulence data and from dozens to

hundreds band positions for molecular markers, whereas a number of isolates tested (sample size) does not exceed 1000 and is usually measured in a few dozens or a couple of hundreds. Then the number of theoretically possible genotypes ($1024 = 2^{10}$ at least) is considerably larger than sample size, which allows normalization of the Shannon entropy by logarithm of sample size.

Similarly the Stoddart diversity index $St(P) = 1/\sum_{r=1}^{s} p_r^2$, which ranges between 1 and $s \leq n$, can be normalized in two different ways: dividing by s and by n, respectively. The first index $St(P)/s$ can be considered as evenness parameter, whereas the second one takes into account both evenness and richness characteristics of populations because

$$\frac{St(P)}{n} = \frac{St(P)}{s} \cdot \frac{s}{n}$$

Genotypic diversity (Shannon, Simpson and Stoddart) and distance (Rogers) indices usually overestimate actual diversity within and difference between populations. This bias becomes more significant for larger number of differentiating characters and can be explained as follows. In the case of k binary characters, the number of theoretically possible genotypes equals 2^k. The number of individuals tested, n, is limited and considerably less than 2^k even for relatively small values of k. Therefore, it is very likely that nearly all sampled individuals are of different genotypes ($s \approx n$ and $p_r \approx 1/n$ for all genotypes $r = 1, 2, \ldots, s$). In such a case, both evenness and richness parameters are very close to their maximum values, which result in high score of genotypic diversity. Actual population diversity might be much lower because of possible similarity between overall allelic patterns of individuals. However, degree of similarity between individuals is not taken into account by indices based only on genotype frequencies. Similarly, the Rogers distance may overestimate actual difference between two populations because of the high probability that none or a very small number of individuals in the two populations will have identical genotypes.

Rationale for choosing diversity indices

What measures of diversity within and distance between populations are suitable for organisms with asexual and mixed mode of reproduction (for instance, plant pathogens)? The choices should take into account that populations with clonal reproduction may differ from each other in ways that sexual random mating populations do not.

Basic property of distance measures All indices of distance between populations should be designed to measure accurately the differences between genetically distinct populations. If any index fails to achieve this main goal in comparisons involving asexual populations, then its applicability for asexual populations must be regarded as having limited value regardless of how well that index measures differences between sexual random mating populations. It is easy to show that the Nei genetic distance, N (Nei, 1972, 1978), mean character difference, MCD (Sneath & Sokal, 1973; Long *et al.*, 1998; Manisterski *et al.*, 2000), Nei's coefficient of differentiation, G_{ST} (Nei, 1973), and the distance of average differences, DAD_m, do not distinguish between two populations P_1 and P_2 that consist of different genotypes but with identical vector of frequencies of occurrence of an assigned value of 1 at each of k differentiating factors (for example, virulence frequencies), i.e. $q_{1i} = q_{2i}$ for all $i = 1, 2, \ldots, k$. The following example illustrates this fact.

Example 2 Consider samples of two individuals each from populations P_1, P_2 and P_3 tested on six differentials.

	P_1							P_2							P_3						
	1	2	3	4	5	6			1	2	3	4	5	6		1	2	3	4	5	6
i_1	1	1	1	0	0	1		i_3	1	1	1	1	0	0	i_5	1	1	0	0	0	1
i_2	1	1	0	1	1	0		i_4	1	1	0	0	1	1	i_6	1	1	1	1	1	0

The vector of virulence frequencies is the same for these samples, and one can calculate that the mean character difference, MCD, the Nei N and G_{ST}, and DAD_m distances between these populations equal 0. Therefore, populations P_1, P_2 and P_3 may be interpreted as 'identical' according to MCD, N, G_{ST} and DAD_m distances, whereas they are composed of different genotypes and therefore distinct. By contrast, the Rogers distance, R, and the Kosman distance, KB_m, do distinguish between these populations: $R(P_1,P_2) = R(P_1,P_3) = R(P_2,P_3) = 1$, $KB_m(P_1,P_2) = 1/3$, $KB_m(P_1,P_3) = 1/6$ and $KB_m(P_2,P_3) = 1/6$. However, these populations are equally and absolutely different according to Rogers' index, which is based on the frequencies of the genotypes, regardless of how many virulences these genotypes share. Conversely, the Kosman distance reveals that populations P_1 and P_3 are more similar than P_1 and P_2. In sexual random mating populations genetic recombination leads to linkage equilibria, because combinations of alleles are continually reshuffled so that natural selection and genetic drift can be said to influence allele frequencies directly. Thus, sexual populations with similar allele frequencies will also have similar genotype frequencies. With asexual (clonal) reproduction, however, natural selection and genetic drift act at the genotype (clone) level rather than directly on allele frequencies. This can generate and maintain linkage disequilibria in which individual genotypes occur at frequencies significantly greater or lower than would be predicted from the frequencies of their component alleles. Therefore, distance measures such as MCD, N, G_{ST}, and DAD_m that are calculated from allele frequencies will nearly always miss a significant part of the genetic differences that occur between asexual populations.

Table 2 Aspects of diversity within populations that are expressed by diversity indices commonly used in population studies

Measures of diversity	Aspects of diversity[a]
Stoddart index (*St*) Simpson index (*Si*) Shannon entropy Shannon normalized index (*Sh*)	(i) Richness (ii) Evenness
Shannon evenness index (*E*)	(ii) Evenness
Average difference (*ADW*) Nei diversity (H_S)	(iii) Differences between genotypes
Kosman diversity (*KW*)	(i) Richness (ii) Evenness (iii) Differences between genotypes (iv) Deviation of differences

[a]Aspects of diversity include the following population parameters: (i) richness = number of genotypes/sample size; (ii) evenness of genotypes distribution; (iii) degree of genetic differences between genotypes; (iv) deviation of pairwise differences between genotypes from the average difference between all genotypes.
Aspects (i) and (ii) may describe species diversity within communities or genotypic diversity within species when it is not feasible to quantify the degrees of dissimilarity between the species or genotypes.
Aspect (iii) alone may sufficiently describe diversity within populations of sexual random mating species.
Aspects (i), (ii), (iii) and (iv) should be considered in describing diversity within species in which asexual reproduction produces clonal lineages and linkage disequilibria.

Basic properties of diversity measures

Plant pathogen populations often exhibit considerable linkage disequilibrium because of asexual or mixed mode of reproduction. An optimal index for measuring diversity within such populations should satisfy several conditions (Groth & Roelfs, 1987). A population is more diverse and diversity index is higher if: (i) that population consists of a larger number of genotypes for a given number of individuals; (ii) it is characterized by a more even distribution of genotypes; (iii) the number of differences in virulence between genotypes within the population is larger.

The Shannon evenness parameter *E* satisfies the property (ii) only (Table 2). The first two properties are fulfilled by Simpson's *Si*, Stoddart's *St* and Shannon's entropy and normalized *Sh* diversity indices (Table 2), but they do not satisfy the third property because the virulence patterns are not taken into account by these indices. By contrast, the measure of average difference within populations *ADW*, the Müller index *Mu* and Nei's measure of the average gene diversity per locus H_S satisfy the property (iii) and do not meet the properties (i) and (ii) (Table 2) because genotypes are not considered by these indices. The Kosman diversity KW_m is a more complicated index. It takes into account the contribution of dissimilarity among individuals to the diversity within a population, not only the relative frequencies of different genotypes like *Si*, *St* and *Sh* indices. In addition, the KW_m index considers a population as a set of different genotypes with possibly associated virulences/avirulences and does not characterize a population by an independent set of virulence frequencies in contrast to *ADW*, *Mu*, H_S and *K* indices. The properties (i) and (ii) are not generally compelling by the Kosman diversity KW_m. The extent of dissimilarity among individuals contributes considerably to the diversity within a population. The following example demonstrates why Kosman's diversity KW_m is preferred over other indices in accounting for multiple aspects of diversity within a population.

Example 3 Consider samples of four individuals each from populations P_1 and P_2 tested on four differentials.

	P_1					P_2			
	1	2	3	4		1	2	3	4
i_1	1	0	0	0	i_5	1	1	0	0
i_2	0	1	0	0	i_6	0	1	0	0
i_3	0	0	1	0	i_7	0	0	0	1
i_4	0	0	0	1	i_8	1	0	0	1

Populations P_1 and P_2 are considered as equally diverse by the following estimators: $Si = 3/4$, $Sh = \ln(4)/\ln(4) = 1$, $ADW_m = H_S = 3/8$, $Mu = 1/2$ are equal for both populations. By contrast, population P_2 is more diverse than P_1 according to Kosman's measure: $KW_m(P_1) = 1/2$ and $KW_m(P_2) = 3/4$.

This example demonstrates that the three properties (i), (ii) and (iii) of an optimal diversity index are not enough to distinguish between population diversities. The number of genotypes in each of two populations P_1 and P_2 is equal (four), the genotypes are evenly distributed within each population (genotype frequencies equal 1/4) and average dissimilarity between genotypes (individuals) equals $ADW_m = 3/8$ for both populations. Nevertheless, the Kosman KW_m index distinguishes between diversities of these populations. It could be explained by the absolute deviation of pairwise dissimilarities from the average dissimilarity:

$$s(P) = \frac{1}{n^2} \cdot \sum_{i,j=1}^{n} \left| m(x_i, x_j) - ADW_m(P) \right| \qquad \text{Eqn 10}$$

(x_i and x_j are individuals from population *P* of *n* individuals). In our case $s(P_1) = 3/16$ is less than $s(P_2) = 1/4$, and this is a reason why population P_2 can be considered as more diverse than population P_1. Therefore, it could be reasonable to supplement three properties of an optimal diversity index by adding the following fourth property. A population is more diverse and diversity index is higher if: (iv) deviation of pairwise differences between genotypes (individuals) from the average difference is larger.

The next example demonstrates that effects of conditions (i)–(iv) on diversity within populations may compensate each other if they counteract.

Example 4 Consider samples of four individuals each from populations P_1 and P_2 tested on four differentials.

	P_1					P_2			
	1	2	3	4		1	2	3	4
i_1	1	0	0	1	i_5	1	1	0	0
i_2	1	1	0	0	i_6	1	1	0	0
i_3	0	1	1	0	i_7	0	0	1	1
i_4	0	0	1	1	i_8	0	0	1	1

The vector of virulence frequencies is the same for the both samples, and therefore, all diversity indices based just on virulence frequencies do not distinguish between diversities within populations P_1 and P_2. On the other hand, according to the Simpson and the Shannon normalized indices, which are based just on genotype frequencies, population P_1 is considered as more diverse than population P_2 ($Si(P_1) = 3/4$, $Si(P_2) = 1/2$) and $Sh(P_1) = \ln(4)/\ln(4) = 1$, $Sh(P_2) = \ln(2)/\ln(4) = 1/2$) because of the larger number of genotypes in P_1 (four) compared with P_2 (two) and even distribution of genotypes in both populations. This means that property (i) was the crucial one for making decision because condition (ii) is satisfied for the both populations P_1 and P_2, and properties (iii) and (iv) are irrelevant for Simpson's and Shannon's indices.

The Kosman diversity KW_m equals 1 for both populations (i.e. P_1 and P_2 are considered as equally diverse). This could be explained as follows. The average differences between isolates are equal for the both populations, $ADW_m = 1/2$, whereas the absolute deviation (Eqn 10) is less for population P_1, $s(P_1) = 1/4$ and $s(P_2) = 1/2$. Thus, populations P_1 and P_2 are of similar characteristics with respect to the conditions (ii) and (iii), but characterizations of these populations by properties (i) and (iv) are of opposite tendency. The greater number of different genotypes in population P_1 increases its level of diversity compared with P_2. However, the measure of variance between isolates is less for population P_1. Therefore, the former effect is compensated by the latter one, which leads to the supposed equal diversity within these populations.

Correlation of results from different types of indices

Different measures of diversity within and distance between populations were applied for analysis of Israeli populations of wheat leaf rust, caused by the fungus *Puccinia triticina* (Manisterski *et al.*, 2000; J. Manisterski, unpublished data). The sexual stage of *P. triticina* occurs on alternate hosts in the genus *Thallictrum*, which do not occur in Israel but do occur in the Iberian peninsula and in Eastern Europe. Thus, *P. triticina* reproduces only asexually within Israel, but some level of migration of *P. triticina* into Israel may occur from sexually reproducing populations in Europe. Dozens of isolates were collected annually during 1993–2002 and tested on a set of 15 differentials, and the corresponding virulence patterns were arranged in two-way data tables of 0s (avirulence) and 1s (virulence). Diversities within 10 annual collections of leaf rust isolates and distances between them were calculated using the KOIND package (Schachtel & Kosman, 2002). The results are presented in the following examples 5 and 6.

Example 5 Comparison of different measures of diversity within populations.

Diversity within 10 populations of leaf rust isolates was estimated by all indices mentioned in Appendix A, *A2*. The corresponding values for all possible pairs of diversity indices were plotted one vs. another using the MXPLOT program of NTSYSpc package, version 2.1 (Exeter Software, Setauket, NY, USA). The results for three diversity measures of different nature (Kosman's diversity within population, *KW*, Nei's measure of the average gene diversity per locus, H_S, and the Shannon normalized index of diversity, *Sh*) are presented in Fig. 1. Rather strong correlation ($r = 0.943$) was observed between values of the Kosman and Nei diversity indices (Fig. 1a). Note, however, that the rank order of least diverse to most diverse was not exactly the same for the Kosman and Nei diversity indices. Results from the Shannon's index were not correlated with those from either the Kosman (Fig. 1b) or Nei (Fig. 1c) diversity indices. Similar results were obtained when the Kosman diversity within population (*KW*) was compared with another Kosman's index (*K*), which is based only on virulence frequencies like the Nei (H_S) index, and with the Simpson (*Si*) and Stoddart (*St*) indices, which are based only on genotype frequencies as is the Shannon normalized index (*Sh*). Weak correlation was observed between an index based only on virulence frequencies (*K* or H_S) and an index based only on genotype frequencies (*Sh*, *Si* or *St*). Conversely, the indices of the same nature were strongly correlated. Thus, different measures can result in qualitatively unlike descriptions of diversity within populations, especially if they are of different nature. As expected for genotypic diversity measures, the Shannon normalized index (*Sh*) is biased towards high values (Fig. 1b,c).

Example 6 Comparison of different measures of distance between populations.

Distance between all possible pairs of 10 populations of leaf rust isolates was estimated by all indices mentioned in Appendix A, A3. The values for all possible pairs of distance indices were plotted one versus another using the MXCOMP program of the NTSYSpc package, version 2.1 (Exeter Software), which also provided the corresponding values of Mantel test correlation between different indices. The results for three distance measures of different nature (Kosman's distance between populations, *KB*, Nei's genetic distance, *N*, and the Rogers distance, *R*) are presented in Fig. 2. Rather strong correlation ($r = 0.951$) was observed between values of the Kosman and Nei distances (Fig. 2a). Note, however, that in some cases the rank order for pairs of populations according

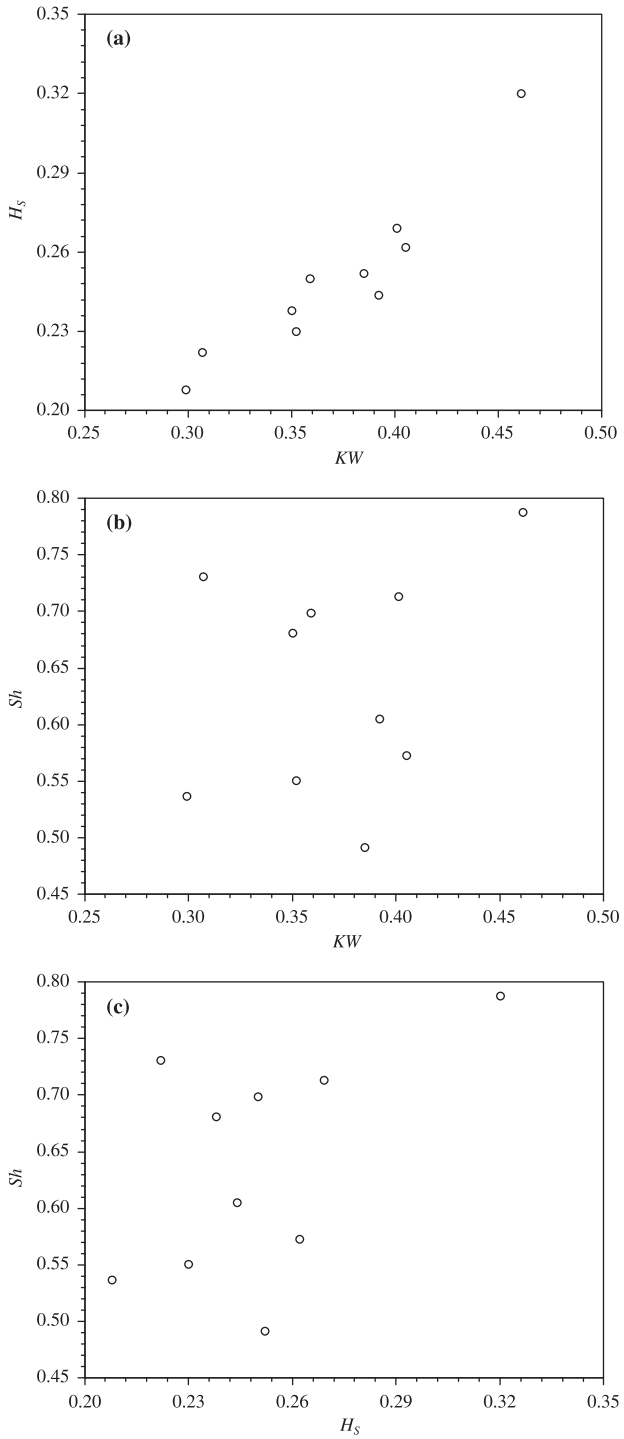

Fig. 1 Comparisons of measures of diversity within 10 annual populations of the wheat leaf rust fungus *Puccinia triticina* collected during the period 1993–2002 and tested for virulence on a set of 15 differentials: (a) the Kosman diversity *KW* vs Nei's measure of the average gene diversity per locus H_S; (b) the Kosman diversity *KW* vs the Shannon normalized index *Sh*; (c) Nei's measure of the average gene diversity per locus H_S vs the Shannon normalized index *Sh*.

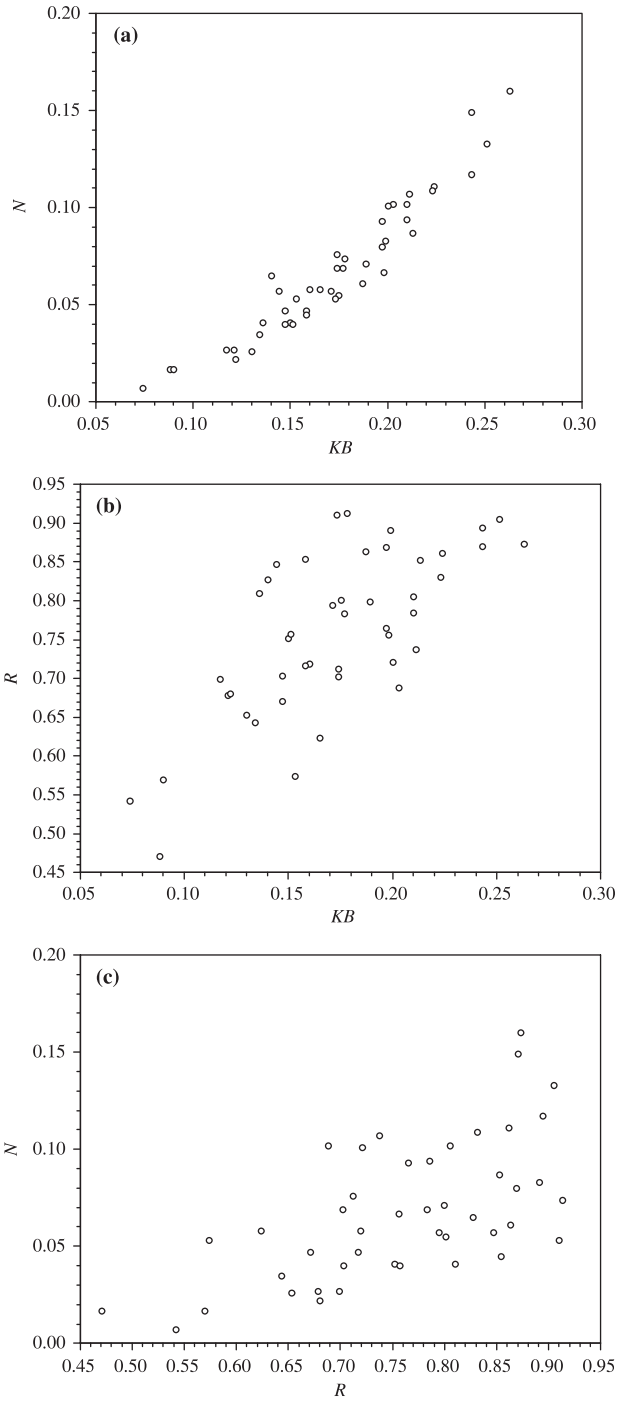

Fig. 2 Comparisons of measures of distance between all possible pairs of 10 annual populations of the wheat leaf rust fungus *Puccinia triticina* collected during 1993–2002 and tested for virulence on a set of 15 differentials: (a) the Kosman distance *KB* vs the Nei genetic distance *N*; (b) the Kosman distance *KB* vs the Rogers distance *R*; (c) the Rogers distance *R* vs the Nei genetic distance *N*.

to distance between them differed considerably between the Kosman and Nei distances. For example, one pair of populations was considered more distant than 23 other pairs by the Nei index, but more distant than only nine other pairs by the Kosman index. A discrepancy of this magnitude could produce a significant distortion in cluster analyses for the populations. Rather strong correlations also were obtained when the Kosman distance (*KB*) was compared with the mean character difference (*MCD*) and Nei's coefficient of differentiation (G_{ST}), which are based only on virulence frequencies as is the Nei distance (*N*). Results from Rogers' distance (*R*), which is based only on genotype frequencies, were only weakly correlated with those from either the Kosman ($r = 0.712$) or the Nei ($r = 0.619$) distances. Weak correlation also was observed between Rogers' distance (*R*) and the group of indices that are based only on virulence frequencies (*MCD*, *N* and G_{ST}). By contrast, the indices of the same nature (*MCD*, *N* and G_{ST}) were strongly correlated. Thus, different measures can result in qualitatively unlike descriptions of distance between populations, especially if they are of different nature. As expected for genotypic indices, the Rogers distance (*R*) is biased towards high values (Fig. 2b,c).

Genetic diversity within and between populations is a function of gene and genotypic structure of populations. Consequently, variance of any genetic population parameter can partly be explained by variability of the corresponding gene and genotypic parameters. The last two are not independent, and it seems that gene structure of populations may have a much stronger impact on genetic estimators than genotypic one has. Therefore, high level of correlation observed between genetic and gene parameters in Examples 5 and 6 (*KW* vs H_S diversity ($r = 0.943$) and *KB* vs *N* distance ($r = 0.951$), respectively) was expected. This means that about 90% (r^2) of the variance of genetic diversity within and between the annual populations of wheat leaf rust can be explained by the variation in corresponding gene estimators. If the *P. triticina* populations in Israel reproduced sexually with random mating, the Kosman and Nei diversities and distances would be even more highly correlated, because the Nei indices assumes linkage equilibrium which typically results from random mating. In asexual populations, however, lack of genetic recombination causes even selectively neutral genes to increase or decrease in frequency in response to average differences in relative fitness of the genotypes in which the neutral alleles happen to occur. Thus, in asexual populations small enough to be affected by genetic drift or in asexual populations exposed to fluctuating environmental conditions from generation to generation, lack of genetic recombination is likely to result in at least some genetic disequilibria even between selectively neutral alleles at different loci.

Example 6 demonstrated that despite strong correlation between distance measures of different types the rank order for pairs of populations according to distance between them may differ considerably. The following example shows that a similar effect may occur even for distance measures of the same type.

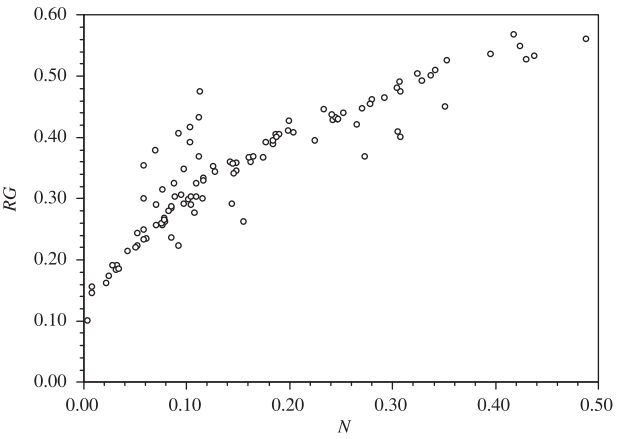

Fig. 3 Comparison of Nei's genetic distance *N* and Rogers' modified genetic distance *RG* between *Apiosporina morbosa* populations isolated from 15 geographic regions (Zhang *et al*., 2005).

Example 7 Comparison of distance measures based on allele frequencies.

Zhang *et al.* (2005) used the sequence-related amplified polymorphism (SRAP) technique to determine genetic diversity and population structure of *Apiosporina morbosa* on *Prunus* spp. Resemblance of the pathogen populations collected from 15 geographic locations was assessed on the basis of allele frequencies at 58 SRAP loci (two alleles corresponding to presence and absence of bands were considered at each putative locus). The Rogers modified genetic distance coefficient (Wright, 1978) and the Nei genetic identity (Nei, 1972) were calculated for all possible pairwise comparisons among 15 populations (Table 3 in Zhang *et al.*, 2005). We calculated the Nei standard genetic distance (*N*) between the populations as a function of Nei identity, and the values of *N* were plotted versus the corresponding values of Rogers modified genetic distance (*RG*) as in Example 6 (Fig. 3). Despite relatively strong correlation ($r = 0.897$) between the indices, in some cases the rank order for pairs of populations according to distance between them differed considerably between the Nei and Rogers modified genetic distances. For example, among total 105 pairs of populations one pair was considered more distant than 93 other pairs by the Rogers genetic distances, but more distant than only 40 other pairs by the Nei index. By contrast, another pair of populations is more distant than only 11 other pairs by the Rogers genetic distances, but more distant than 34 other pairs by the Nei index. A discrepancy of this magnitude could produce a significant distortion in cluster analyses for the populations.

Conclusions

We compared different methods of diversity analysis of populations with asexual or mixed mode of reproduction (Table 1). Relationships between the diversity indices of different types are generally unpredictable, and different measures of

diversity within and distance between populations may result in qualitatively different inferences. Even if correlation between distance indices is strong, the rank order for some pairs of populations according to distance between them may differ considerably, which could affect results of cluster analysis. Although we considered only examples with plant pathogens, our results are relevant for other organisms, including higher plants that have an asexual or mixed mode of reproduction.

Kosman's distance between populations and diversity within population (Eqn 5) are the only diversity indices that take into account both the genotypic structure of populations and measure of similarity between different genotypes. Therefore, they are more suitable and informative than gene diversity and distance indices based only on gene frequencies for analysing populations that exhibit a significant degree of linkage disequilibrium. The Kosman distances with respect to the simple mismatch coefficient and the Jaccard dissimilarity are metrics. This means that they always yield a positive value for difference between distinct populations and satisfy the triangle inequality (see Appendix B), which may be important for some clustering procedures. The Shannon, Simpson and Stoddart diversity indices and Rogers distance are appropriate for comparisons of species diversity within or between ecological units when it is not feasible to quantify degrees of difference between species. These measures are also suitable when only genotypic structure of populations is studied. However, genotypic diversity and distance measures may be excessively biased towards high values when comparing with the corresponding genetic and gene parameters.

Nei's measure of the average gene diversity per locus (Nei, 1973), the index of average difference within population (Eqn 1) with respect to the simple mismatch dissimilarity (McCain *et al*., 1992) and the Müller diversity index (Müller *et al*., 1996) are in fact the same measure of diversity within populations (Kosman, 2003). The distance of average differences (Eqn 2) with respect to the simple mismatch coefficient equals the Nei minimum genetic distance (Nei, 1972). The Nei standard genetic distance (Nei, 1972, 1978) is a function of the indices of average difference within and between populations with respect to the simple mismatch coefficient (Eqn 3). All of these measures, which are based on allele frequencies or can be derived from allele frequencies as well as from average differences between genotypes, are suitable measures of diversity in populations of species with sexual random mating. For reasons described herein, however, these measures generally provide an incomplete representation of genetic diversity in species with asexual or mixed mode of reproduction. The Kosman indices of distance between populations and diversity within populations are designed to overcome this limitation.

Note

Software for calculating Kosman's diversity indices and some other population parameters is free and available from the first author.

Acknowledgements

We are grateful to J. Manisterski (Tel Aviv University, Israel) for the data placed at our disposal. This work was partially supported by the Leiberman-Okinow and Colton foundations (Tel Aviv University), and GIF Research Grant No. I-744-121.12/2002 (German–Israeli Foundation for Scientific Research and Development).

References

Adhikari TB, Mew TW, Leach JE. 1999. Genotypic and pathotypic diversity in *Xanthomonas oryzae* pv. *Oryzae* in Nepal. *Phytopathology* **89**: 687–694.

Andrivon D, de Vallavieille-Pope C. 1995. Race diversity and complexity in selected populations of fungal biotrophic pathogens of cereals. *Phytopathology* **85**: 897–905.

Gale LR, Chen LF, Hernick CA, Takamura K, Kistler HC. 2002. Population analysis of *Fusarium graminearum* from wheat fields in eastern China. *Phytopathology* **92**: 1315–1322.

Goodwin SB, Spielman LJ, Matuszak JM, Bergeron SN, Fry WE. 1992. Clonal diversity and genetic differentiation of *Phytophthora infestans* populations in northern and central Mexico. *Phytopathology* **82**: 955–961.

Groth JV, Roelfs AP. 1987. The concept and measurement of phenotypic diversity in *Puccinia graminis* on wheat. *Phytopathology* **77**: 1395–1399.

Grünwald NJ, Goodwin SB, Milgroom MG, Fry WE. 2003. Analysis of genotypic diversity data for populations of microorganisms. *Phytopathology* **93**: 738–746.

Kolmer JA, Liu JQ. 2000. Virulence and molecular polymorphism in international collections of the wheat leaf rust fungus *Puccinia triticina*. *Phytopathology* **90**: 427–436.

Kosman E. 1996. Difference and diversity of plant pathogen populations: a new approach for measuring. *Phytopathology* **86**: 1152–1155.

Kosman E. 2003. Nei's gene diversity and the index of average differences are identical measures of diversity within populations. *Plant Pathology* **52**: 533–535.

Kosman E, Leonard KJ. 2005. Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology* **14**: 415–424.

Long DL, Leonard KJ, Roberts JJ. 1998. Virulence and diversity of wheat leaf rust in the United States in 1993–95. *Plant Disease* **82**: 1391–1400.

Lynch M. 1990. The similarity index and DNA fingerprinting. *Molecular Biology and Evolution* **7**: 478–484.

Lynch M, Milligan BG. 1994. Analysis of population genetic structure with RAPD markers. *Molecular Ecology* **3**: 91–99.

Manisterski J, Eyal Z, Ben-Yehuda P, Kosman E. 2000. Comparative analysis of indices in the study of virulence diversity between and within populations of *Puccinia recondita* f. sp. *tritici* in Israel. *Phytopathology* **90**: 601–607.

McCain JW, Groth JV, Roelfs AP. 1992. Inter- and intrapopulation isozyme variation in collections from sexually reproducing populations of the bean rust fungus, *Uromyces appendiculatus*. *Mycologia* **84**: 329–340.

Müller K, McDermott JM, Wolfe MS, Limpert E. 1996. Analysis of diversity in populations of plant pathogens: the barley powdery mildew pathogen across Europe. *European Journal of Plant Pathology* **102**: 385–395.

Nei M. 1972. Genetic distance between populations. *American Naturalist* **106**: 283–292.

Nei M. 1973. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences, USA* **70**: 3321–3323.

Nei M. 1978. Estimation of average heterozygosities and genetic distance from a small number of individuals. *Genetics* **89**: 583–590.

Rogers JS. 1972. Measures of genetic similarity and genetic distance, pp. 145–143. In: *Studies in genetics*. Austin, TX, USA: University of Texas.

Schachtel GA, Kosman E. 2002. KOIND package – Short manual. In: *Online publication biometrie and populationsgenetic, JLU Giessen.* Accessed at http://www.va-tipp.de.

Shannon CE, Weaver W. 1949. *The mathematical theory of communication*. Urbana, IL, USA: University of Illinois Press.

Sheldon AL. 1969. Equitability indices: Dependence on the species count. *Ecology* **50**: 466–467.

Simpson EH. 1949. Measurement of diversity. *Nature* **163**: 688.

Sneath PA, Sokal RR. 1973. *Numerical taxonomy*. San Francisco, CA, USA: W.H. Freeman.

Stoddart JA. 1983. A genotypic diversity measure. *Journal of Heredity* **74**: 489–490.

Stoddart JA, Taylor JF. 1988. Genotypic diversity: estimation and prediction in samples. *Genetics* **118**: 705–711.

Wright S. 1978. *Evolution and the genetics of populations*, Vol. 4. *Variability within and among natural populations.* Chicago, IL, USA: University of Chicago Press.

Zhang JX, Fernando WGD, Remphrey WR. 2005. Genetic diversity and structure of the *Apiosporina morbosa* populations on *Prunus* spp. *Phytopathology* **95**: 859–866.

Appendix A

A1 Consider binary response patterns of two individuals x_1 and x_2 on a set of k factors. Denote a = number of factors with positive response for the both individuals, b = number of factors where x_1 and x_2 respond positively and negatively, respectively, and c = number of factors where x_1 and x_2 respond negatively and positively, respectively.

Measures of similarity/dissimilarity between individuals

Coefficient	Symbol	Formula
Jaccard coefficient of similarity	J	$J(x_1,x_2) = a/(a + b + c)$
Jaccard coefficient of dissimilarity	j	$j(x_1,x_2) = 1 - J(x_1,x_2)$
Dice coefficient of similarity	D	$D(x_1,x_2) = 2a/(2a + b + c)$
Dice coefficient of dissimilarity	d	$d(x_1,x_2) = 1 - D(x_1,x_2)$
Simple match coefficient	M	$M(x_1,x_2) = (k - b - c)/k$
Simple mismatch coefficient	m	$m(x_1,x_2) = 1 - M(x_1,x_2)$
Euclidean distance	e	$e(x_1,x_2) = (b + c)^{1/2}$

A2 Consider a sample collected from population P, which consists of n individuals $x_1, x_2, \ldots, x_n$ tested on k differentiating factors and represented by binary patterns. We denote by q_i the frequency of appearance 1 at the ith differentiating factor, $i = 1, 2, \ldots, k$. The number of individuals and frequency of genotype r in population P are denoted by n_r and p_r, $r = 1, 2, \ldots, s$, respectively, where s is the number of different genotypes in P. The measure of dissimilarity between individuals is denoted by ρ (see A1, where different measures $\rho = j, d, m$ or e are explained).

Measures of diversity within population

Index	Symbol	Formula/reference
Average distance within	ADW	$ADW_\rho(P) = \sum \rho(x_i,x_j)/n^2,\ 1 \le i, j \le n$
Kosman diversity within	KW	$KW_\rho(P) = Ass^\rho_{max}(P,P)/n$
Kosman index	K	$K(P) = \sum \min[2q_i, 2(1 - q_i)]/k,\ 1 \le i \le k$
Müller index of diversity	Mu	$Mu(P) = \sum m(x_i,x_j)/[n(n - 1)/2],\ 1 \le i < j \le n$ (Müller *et al*., 1996)
		$Mu(P) = [2n/(n - 1)]\sum q_i(1 - q_i)/k,\ 1 \le i \le k$ (Manisterski *et al*., 2000)
Nei measure of gene diversity	H_S	$H_S(P) = \sum[1 - q_i^2 - (1 - q_i)^2]/k,\ 1 \le i \le k$
Simpson index	Si	$Si(P) = 1 - \sum p_r^2,\ 1 \le r \le s$
Stoddart index	St	$St(P) = 1/\sum p_r^2,\ 1 \le r \le s$
Shannon evenness parameter	E	$E(P) = -(\sum p_r \ln p_r)/\ln s,\ 1 \le r \le s$
Shannon normalized index	Sh	$Sh(P) = -(\sum p_r \ln p_r)/\ln n,\ 1 \le r \le s$

A3 Consider two samples collected from two populations P_1 and P_2, which consist of the same number n of individuals x_{11}, x_{12}, … , x_{1n} and x_{21}, x_{22}, … , x_{2n}, respectively, tested on k differentiating factors and represented by binary patterns. We denote by q_{1i} and q_{2i} the frequencies of appearance 1 at the ith differentiating factor for populations P_1 and P_2, respectively. The frequencies of genotype r in populations P_1 and P_2 are denoted by p_{1r} and p_{2r}, respectively, $r = 1, 2, \dots, s$, where s is the total number of different genotypes in both populations. The measure of dissimilarity between individuals is denoted by ρ (see A1, where different measures $\rho = j, d, m$ or e are explained).

Measures of diversity between populations

Index	Symbol	Formula/reference
Average distance between	*ADB*	$ADB_\rho(P_1,P_2) = \sum \rho(x_{1i},x_{2j})/n^2,\ 1 \le i, j \le n$
Distance of average differences	*DAD*	$DAD_\rho(P_1,P_2) = ADB_\rho(P_1,P_2) - [ADW_\rho(P_1) + ADW_\rho(P_2)]/2$
Kosman distance	*KB*	$KB_\rho(P_1,P_2) = Ass^\rho_{min}(P_1,P_2)/n$
Mean character difference	*MCD*	$MCD(P_1,P_2) = \sum\lvert q_{1i} - q_{2i}\rvert/k,\ 1 \le i \le k$
Nei genetic distance	*N*	* (Nei, 1972, 1978)
Nei minimum genetic distance	N_M	* (Nei, 1972)
Nei coefficient of differentiation	G_{ST}	* (Nei, 1973)
Rogers distance	*R*	$R(P_1,P_2) = \sum\lvert p_{1r} - p_{2r}\rvert/2,\ 1 \le r \le s$

*Refer to reference for explanations and formula.

Appendix B

Let us consider any set X and any nonnegative function ρ determined for each ordered pair of elements from X. This function ρ is called metric if the following properties are fulfilled for all elements x_1, x_2 and x_3 from X:

1 $\rho(x_1,x_2) \ge 0$, and $\rho(x_1,x_2) = 0$ if and only if $x_1 = x_2$.
2 $\rho(x_1,x_2) = \rho(x_2,x_1)$.
3 The triangle inequality $\rho(x_1,x_2) + \rho(x_2,x_3) \ge \rho(x_1,x_3)$.

The function ρ is called pseudo-metric if property 1 is replaced by

1a. $\rho(x_1,x_2) \ge 0$, and $x_1 = x_2$ implies $\rho(x_1,x_2) = 0$.

Therefore, in the case of pseudo-metric it is possible that $x_1 \ne x_2$ but $\rho(x_1,x_2) = 0$.

